European Radiology
https://doi.org/10.1007/s00330-021-08298-7

ULTRASOUND

An integrated AI model to improve diagnostic accuracy of ultrasound and output known risk features in suspicious thyroid nodules

Juan Wang[1] · Jue Jiang[1] · Dong Zhang[2] · Yao-zhong Zhang[3] · Long Guo[4] · Yusheng Jiang[5] · Shaoyi Du[6] · Qi Zhou[1]

Received: 2 April 2021 / Revised: 16 August 2021 / Accepted: 23 August 2021

Abstract

Objectives From the viewpoint of ultrasound (US) physicians, an ideal thyroid US computer-assisted diagnostic (CAD) system for thyroid cancer should perform well in suspicious thyroid nodules with atypical risk features and be able to output explainable results. This study aims to develop an explainable US CAD model for suspicious thyroid nodules.

Methods A total of 2992 solid or almost-solid thyroid nodules were analyzed retrospectively. All nodules had pathological results (1070 malignancies and 1992 benignities) confirmed by ultrasound-guided fine-needle aspiration cytology and histopathology after thyroidectomy. A deep learning model (ResNet50) and a multiple risk features learning ensemble model (XGBoost) were used to train the US images of 2794 thyroid nodules. Then, an integrated AI model was generated by combining both models. The diagnostic accuracies of the three AI models (ResNet50, XGBoost, and the integrated model) were predicted in a testing set including 198 thyroid nodules and compared to the diagnostic efficacy of five ultrasonographers.

Results The accuracy of the integrated model was 76.77%, while the mean accuracy of the ultrasonographers was 68.38%. Of the risk features, microcalcifications showed the highest contribution to the diagnosis of malignant nodules.

Conclusions The integrated AI model in our study can improve the diagnostic accuracy of suspicious thyroid nodules and output the known risk features simultaneously, thus aiding in training young ultrasonographers by linking the explainable results to their clinical experience and advancing the acceptance of AI diagnosis for thyroid cancer in clinical practice.

Key Points

- *We developed an artificial intelligence (AI) diagnosis model based on both deep learning and multiple risk feature ensemble learning methods.*
- *The AI diagnosis model showed higher diagnostic accuracy for suspicious thyroid nodules than ultrasonographers.*
- *The AI diagnosis model showed partial explainability by outputting the known risk features, thus aiding young ultrasonic doctors in increasing the diagnostic level for thyroid cancer.*

Keywords Ultrasound · Thyroid cancer · Thyroid nodules · Artificial intelligence · Deep learning

Abbreviations

ACR	The American College of Radiology
AI	Artificial intelligence
AUC	The area under the curve

Juan Wang and Jue Jiang contributed equally to this work and are considered to be co-first authors.

✉ Shaoyi Du
dushaoyi@gmail.com

✉ Qi Zhou
usai_2020@163.com

[1] Department of Ultrasound, the Second Affiliated Hospital, Medical School of Xi'an Jiaotong University, Xi'an 710004, China

[2] School of Automation Science and Engineering, Faculty of Electronic and Information Engineering, Xi'an Jiaotong University, Xi'an 710049, Shaanxi, China

[3] The Institute of Medical Science, The University of Tokyo, Shirokanedai 4-6-1, Minato-ku, Tokyo 108-8639, Japan

[4] Precision Medicine Institute, Western China Science and Technology Innovation Harbor, Xi'an Jiaotong University, Xi'an 712000, China

[5] Department of Computer Science, Columbia University, 500 West 120 Street, New York, NY 10027, USA

[6] Present Address: Institute of Artificial Intelligence and Robotics, College of Artificial Intelligence, Xi'an Jiaotong University, Xi'an 710049, China

Published online: 18 October 2021

DL	Deep learning
FNA	Fine-needle aspiration
ROI	Region of interest
TI-RADS	Thyroid Image Radiology and Data System
US CAD	US computer-assisted diagnosis
US	Ultrasound

Introduction

Thyroid cancer is the most common malignant tumor in the endocrine system [1], and its prevalence has been increasing recently [2]. Ultrasound (US) imaging is a primary tool to find thyroid nodules and differentiate thyroid cancer from benign nodules. However, the known US grayscale risk features to predict malignant thyroid nodules, including solid composition, hypoechogenicity, irregular margin, microcalcifications, and taller-than-wide shape [3], overlap between benign and malignant nodules, thus making the differential diagnosis difficult. Additionally, US examination is strongly operator-dependent, and the differential diagnosis accuracy shows great interobserver variation [4]. These factors tend to cause misdiagnosis, leading to unnecessary fine-needle aspiration (FNA) and thyroidectomy, subsequently resulting in a poor quality of life [5].

To guide clinical decision-making properly, the American College of Radiology (ACR) developed a system of risk classification based on the comprehensive scores of nodule risk features, named the Thyroid Image Radiology and Data System (TI-RADS), which is used to guide whether the patient should undergo FNA cytology or follow-up [6]. However, suspicious thyroid nodules usually classified as TR4 and TR5 show quite a wide range (> 5%) of malignancy probability. To increase the diagnostic accuracy of these suspicious thyroid nodules, it is imperative to determine how important these known risk features are and to further explore new risk features.

Recently, artificial intelligence (AI), especially deep learning (DL) methods, has shown a powerful self-learning ability to extract multilayer and complex features from lesion images independent of ultrasonographers. Several thyroid US computer-assisted diagnosis (US CAD) models using DL methods have been developed to aid ultrasonographers in diagnosing thyroid nodules [7–10]. These models are highly sensitive in finding nodules from thyroid US images and show high accuracy in differentiating malignant nodules from benign nodules. However, most reported models exhibited similar specificity or a small increase in specificity without statistical significance when comparing highly experienced physicians [7, 11]. As physicians know, in comparison to thyroid nodules of TR1-3, it is much more difficult to distinguish the malignant and benign nodules of TR4-5 due to their overlapping risk features between the benign and malignant nodules. Thus, it is reasonable to postulate that the diagnostic specificity of the US CAD models may decrease when the proportion of suspicious thyroid nodules increases. Actually, for both AI and humans, it is difficult to exactly assess the risk level of suspicious nodules with atypical features. Moreover, from the perspective of ultrasonographers, an excellent AI diagnosis system will let them know what risk features are learned and used for diagnosing malignant nodules rather than just outputting the final result, which is the so-called black box deficiency in DL methods [12, 13]. This may be one of the reasons that ultrasonographers tend to have negative attitudes toward the clinical application of US CAD models. Thus, an ideal AI thyroid diagnosis model should have better performance in diagnosing suspicious thyroid nodules than rich-experienced ultrasonographers and simultaneously provide some physician-interpretable evidence to support the diagnosis. In this study, we retrospectively develop an explainable US CAD model for suspicious thyroid nodules based on a training dataset of 2794 and a testing dataset of 198 US images of thyroid nodules.

Materials and methods

Study population

This retrospective study was Health Insurance Portability and Accountability Act–compliant, approved by the institutional review board (No. 2018200), and included patients from a single academic medical center. Informed consent was obtained from all the subjects. The initial population included 3892 thyroid nodules with solid or predominantly solid composition in 3810 patients who underwent thyroid US, US-guided FNA, or thyroidectomy between December 2014 and December 2017 in the Second Affiliated Hospital of Xi'an Jiaotong University. We excluded 475 nodules in 475 patients since they failed to obtain the final pathological diagnoses by FNA. Then, 73 thyroid images from 73 patients with more than one nodule were also excluded. We also excluded 352 thyroid nodules of US images with poor quality. Finally, 2992 nodules (including 835 TR4 and 2157 TR5) from 2992 patients were analyzed in this study. Of these nodules, 2794 nodules (including 743 TR4 and 2051 TR5) were randomly assigned to the training set, and the remaining 198 nodules (92 TR4 and 106 TR5) were assigned to the testing set (Fig. 1). All thyroid nodule images were obtained from the same device (HITACHI HI VISION Preirus) using an EUP-L75 probe at 5–18 MHz. The operation and diagnosis of the nodules were independently accomplished by two ultrasonographers with 26 years and 17 years of experience in thyroid US diagnosis. All the selected thyroid nodules were scored as TR4 and TR5 adhering to ACR TI-RADS[6] based on saved static images.

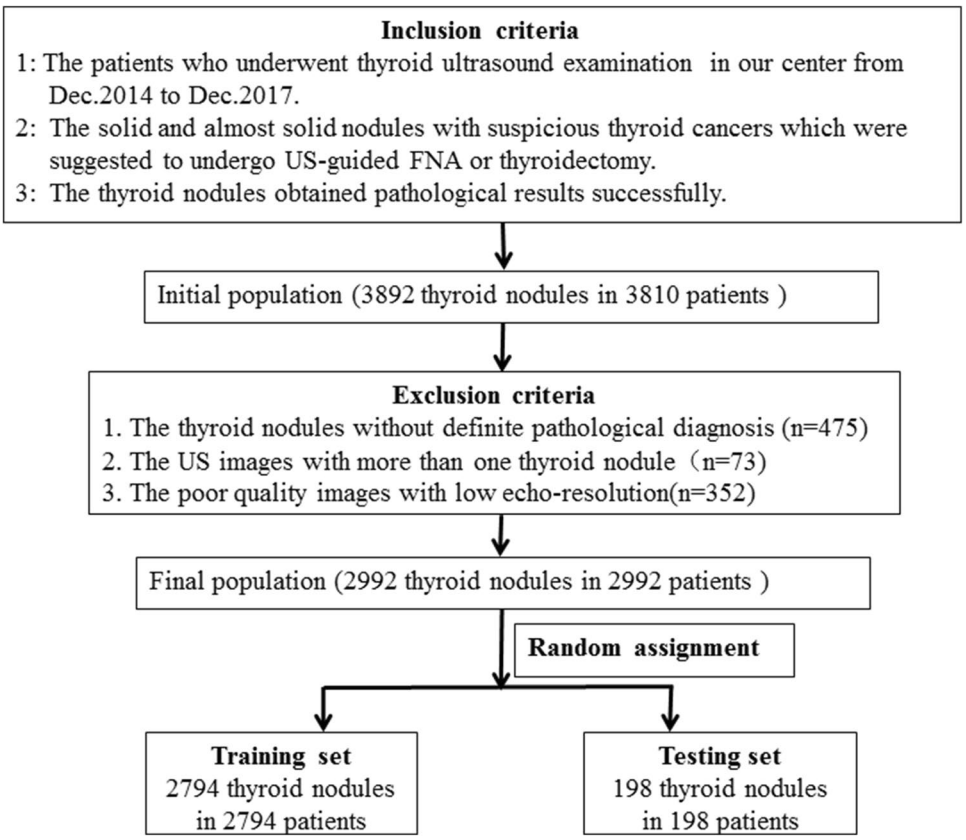

Fig. 1 Inclusion criteria flowchart for the initial population and exclusion criteria for the final study population. US, ultrasound, FNA, fine-needle aspiration

The cytopathological and/or histopathological results were obtained by US-guided FNA and/or surgical resection.

Algorithm development of US CAD models

The thyroid grayscale ultrasonic images were preprocessed. Briefly, the boundaries of all the nodules in the images were manually outlined by 11 ultrasonographers who were trained to recognize nodules from the thyroid tissue of US images successfully. Then, the nodules and their surrounding region of interest (ROI) were extracted and saved automatically by the adaptive cropping method. In the training process, a deep learning method and a traditional feature learning method were applied separately to build the US CAD models based on the training dataset of 2794 thyroid nodule US images. To optimize the model for predicting the probability of malignant nodules, we first built an initial model using ResNet50 [14] based on the ImageNet dataset [15]. Then, we fine-tuned the parameters of this model by a transfer learning method based on our training dataset.

The known US grayscale risk features for predicting malignant thyroid nodules include solid composition, hypoechogenicity, microcalcifications, and taller-than-wide shape [6], which are adopted by the ACR TI-RADS. The taller-than-wide shape is defined as anteroposterior diameter longer than the transverse or longitudinal diameter on the transverse or longitudinal image, known to be less sensitive, but highly specific for malignancy, with a specificity of 88.4–98.7% [16]. Microcalcification is defined as echogenic foci of 1 mm or less with or without posterior acoustic shadowing within the solid portion and is reported to be highly suggestive of malignancy, with a reported specificity of 84–97% [16]. To explore the importance of these risk features that concern ultrasonographers, we first built a traditional machine learning model for recognizing and extracting different types of US features automatically by defining their mathematical functions. These US features included shape (the ratio between anteroposterior and transverse diameter), composition (solid or mixed), and echogenicity (hypoechogenicity or not). In contrast, the calcification features (microcalcification or not) were automatically learned and extracted by a U-net model due to its complexity. Finally, all the models of feature learning were combined into a single model to predict the probability of malignant nodules using an ensemble method (XGBoost) [17], by which the importance of the risk features could be weighted.

Finally, we integrated the predictive probability from the ResNet50 and XGBoost models into a new model. The training process is shown in Fig. 2, and the codes of the CAD algorithms were available on the link (https://github.com/martj001/ai_thyroid).

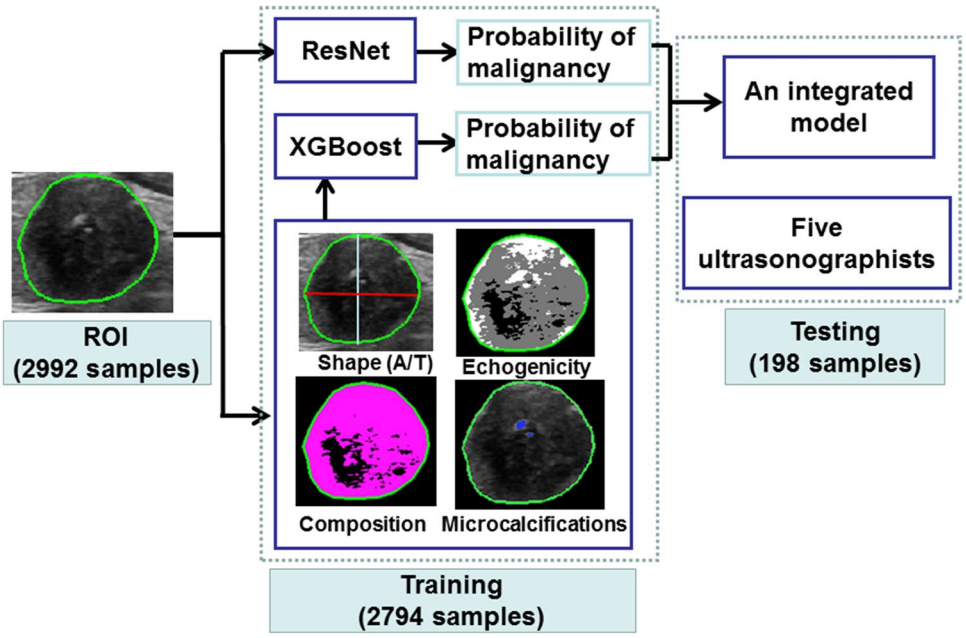

Fig. 2 Training and testing summary of the artificial intelligence ultrasonic diagnosis models for thyroid nodules

Comparing the performance of AI US CAD models and ultrasonographers

US images from 198 thyroid nodules were used as a testing set for assessing the performances of the ResNet50 model, the XGBoost model, and the integrated model. Five ultrasonographers, including two seniors (10 years and 17 years of working experience) and three juniors (2 years, 3 years, and 5 years of working experience), annotated the risk features of each nodule and diagnosed the nodules as malignant or benign independently. The five ultrasonographers were not included in the above mentioned 11 ultrasonographers who were trained to find nodules from the thyroid US images and were not in charge of deciding whether the nodules were malignant. All five ultrasonographers were blinded to the histopathological results. The area under the curve (AUC) of the receiver operating characteristic (ROC) curve was used to compare the diagnostic accuracy among the models and the ultrasonographers.

Statistical analysis

In this study, the AUC of the ROC curve, sensitivity, and specificity were calculated for the ResNet50 model, the XGBoost model, the integrated model and the diagnostic results of ultrasonographers by using a test for differences between malignant and benign thyroid nodules. The chi-square tests and t-tests were used to compare the differences in patient age and sex between benign and malignant lesions separately for the training and testing sets. The McNemar test was used to compare the differences in sensitivity, specificity and accuracy among AI models and ultrasonographers. Statistical analysis was conducted by using R software (R Foundation for Statistical Computing; https://r-project.org). The codes for the analysis are shown in the supplementary material.

Results

Participant demographics

We analyzed 2992 grayscale ultrasonic images of suspicious thyroid nodules from 2992 patients who received US and histopathological examination. The histopathological results showed 1922 benign nodules (64.2%) and 1070 malignant nodules (35.8%) (Table 1). A total of 2,382 patients (79.7%) were female, with a mean age of 50.2 ranging from 13 to 90 years, while 607 patients (20.3%) were male, with a mean age of 46.4 ranging from 10 to 88 years. The age of three patients was unknown. Of 2,794 nodules in the training set, 992 (35.5%) were malignant nodules from 992 patients (203 males and 788 females) with a mean age of 46.4 years, while 1802 (64.5%) were benign nodules from 1802 patients (367 males and 1434 females) with a mean age of 50.1 years (Table 1). The types of US features, including taller-than-wide (A/T) shape, composition, echogenicity, and calcification, were significantly different between malignant and benign nodules ($p < 0.001$). Of 198 nodules in the testing set, 78 (39.4%) were malignant nodules from 78 patients (13 males and 65 females) with a mean age of 47.0 years, while 120 (66.6%) were benign nodules from 120 patients (24 males and 95 females) with a mean age of 50.8 years (Table 1). Except for solid composition ($p = 0.19$), other risk

Table 1 Patient demographics and US features in the training and testing sets

Demographic	Training set		Testing set	
	Benign ($n=1802$)	Malignant ($n=992$)	Benign ($n=120$)	Malignant ($n=78$)
Sex*				
Female	1434 (79.6)	788 (79.4)	95 (79.2)	65 (83.3)
Male	367 (20.4)	203 (20.5)	24 (20.0)	13 (16.7)
Unknown	1	1	1	0
Age (years)[#]				
All	50.1 ± 12.8	46.4 ± 12.1	50.8 ± 13.0	47.0 ± 12.0
Female	49.9 ± 12.5	46.1 ± 11.8	50.7 ± 12.5	47.0 ± 11.3
Male	47.3 ± 13.2	47.5 ± 13.2	46.7 ± 15.9	46.7 ± 15.9
Unknown	3	0	0	0
Shape (A/T)*				
> 1	252 (14.0)	729 (73.5)	103 (85.8)	57 (73.1)
< 1	1550 (86.0)	263 (26.5)	17 (14.2)	21 (26.9)
Composition*				
Solid (area > 95%)	1274 (70.7)	602 (60.7)	79 (65.8)	42 (53.9)
Mixed	528 (29.3)	390 (39.3)	41 (34.2)	36 (46.2)
Cystic (area > 95%)	0	0	0	0
Echogenicity*				
Hypoechogenicity	1374 (76.3)	872 (88.0)	98 (81.7)	72 (92.3)
Iso/hyperechogenicity, mixed echogenicity	428 (23.8)	120 (12.1)	22 (18.3)	6 (7.7)
Calcification*				
Microcalcifications	523 (29.0)	496 (50.0)	86 (71.7)	44 (56.4)
Absence or macrocalcifications	1279 (71.0)	496 (50.0)	34 (28.3)	34 (43.6)

US, ultrasound; *Number of patients. Numbers in parentheses represent percentage; [#]Data are means ± standard deviation. *A/T*, anteroposterior/transverse diameter

features were significantly different between malignant and benign nodules ($p < 0.001$).

Diagnostic performance of ResNet50 and XGBoost US CAD models

The diagnostic performance of the ResNet50 and XGBoost US CAD models in the testing set was evaluated by calculating the AUC, accuracy, sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV) (Fig. 3, Table 2). As a result, the ResNet50 model showed an accuracy of 75.25%, a sensitivity of 74.36%, a specificity of 75.83%, a PPV of 66.67%, an NPV of 81.98%, and an AUC of 0.79 with a cutoff value of 0.43. The XGBoost model showed an accuracy of 73.23%, a sensitivity of 64.10%, a specificity of 79.71%, a PPV of 66.67%, an NPV of 77.24%, and an AUC of 0.77 with a cutoff value of 0.48. Both of the CAD models showed statistically significant differences in differentiating malignant from benign thyroid nodules ($p < 0.05$).

Different from the ResNet50 model, which only showed the diagnostic result (malignancy or benignancy), the XGBoost model not only displayed the diagnostic result but also indicated the existence or nonexistence of the known US grayscale risk features of thyroid cancer in each nodule (Fig. 4). Among these risk features, microcalcifications showed the highest contribution to malignant nodules, with a weighting coefficient of 0.55 (Fig. 5). The feature with the second-highest contribution was a taller-than-wide shape (A/T > 1) with a coefficient of 0.24, followed by hypoechogenicity and solid composition with coefficients of 0.13 and 0.08, respectively (Fig. 5).

Performance comparison among US CAD models and ultrasonographers

We combined the predictive probability from the ResNet50 and XGBoost models to generate an integrated model (Fig. 2). The integrated model showed a diagnostic accuracy of 76.77%, a sensitivity of 69.23%, a specificity of 81.67%, a PPV of 71.05%, an NPV of 80.33%, and an AUC of 0.80 with a cutoff value of 0.47, where the specificity of the integrated model was significantly higher than ResNet model ($p < 0.05$); meanwhile, the AUC, sensitivity, and accuracy did not show significant differences when compared to the AI models alone ($p > 0.05$) (Fig. 3,

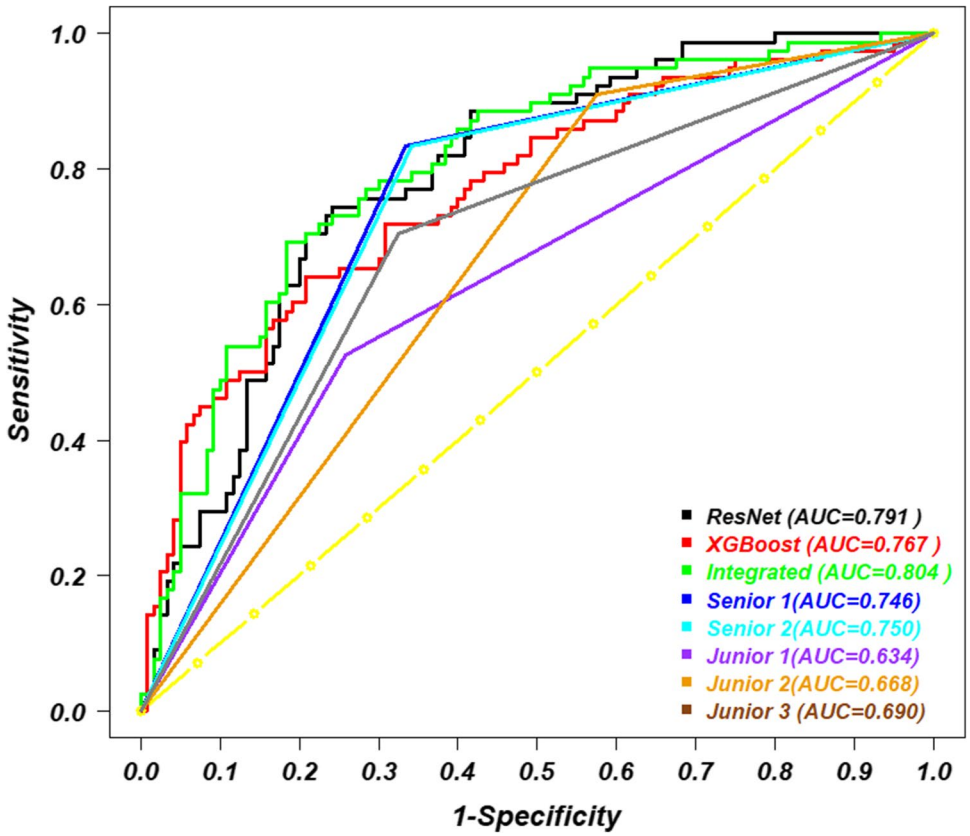

Fig. 3 Diagnostic performance comparison among artificial intelligence models and ultrasound physicians

Table 2 Performance comparison among ultrasonographers and computer-assisted diagnosis models

Diagnosis marker	ROC AUC (95% CI)	Cutoff value	SEN (%)	SPE (%)	ACC (%)	PPV (%)	NPV (%)
Ultrasonographers (working year)							
Junior 1 (2)	0.63 (0.57,0.70)	0.50	41/78(52.56)	89/120(74.17)	130/198(65.66)	41/72(56.94)	89/126(70.63)
Junior 2 (3)	0.67 (0.61,0.72)	0.50	71/78(91.03)	51/120(42.50)	122/198(61.62)	71/140(50.71)	51/58(87.93)
Junior 3 (5)	0.69 (0.62,0.76)	0.50	55/78(70.51)	81/120(67.50)	136/198(68.69)	55/94(58.51)	81/104(77.88)
Mean of Juniors	/	/	(71.37)	61.39	(65.32)	(55.39)	(78.81)
Senior 1 (10)	0.75 (0.69,0.81)	0.50	65/78(83.33)	79/120(65.83)	144/198(72.72)	65/106(61.32)	79/92(85.87)
Senior 2 (17)	0.75 (0.69,0.81)	0.50	65/78(83.33)	80/120(66.67)	145/198(73.23)	65/105(61.90)	80/93(86.02)
Mean of seniors	/	/	(83.33)	(66.25)	(72.98)	(61.61)	(85.95)
Mean of sonographers	/	/	(76.15)	(63.33)	(68.38)	(59.90)	(81.68)
AI models							
XGBoost	0.77 (0.70,0.84)	0.48	50/78(64.10)	95/120(79.17)	145/198(73.23)	50/75(66.67)	95/123(77.24)
ResNet50	0.79 (0.73,0.85)	0.43	58/78(74.36)	91/120(75.83)	149/198(75.25)	58/87(66.67)	91/111(81.98)
Integrated	0.80 (0.74,0.87)	0.47	54/78(69.23)	98/120(81.67)	152/198(76.77)	54/76(71.05)	98/122(80.33)

Working year in the parentheses represents the working experience of US diagnosis for thyroid modules. *SEN*, sensitivity; *SPE*, specificity; *PPV*, positive predictive value; *NPV*, negative predictive value; *ROC*, receiver operating characteristic curve; *AUC*, area under the receiver operating curve; *CI*, confidence interval

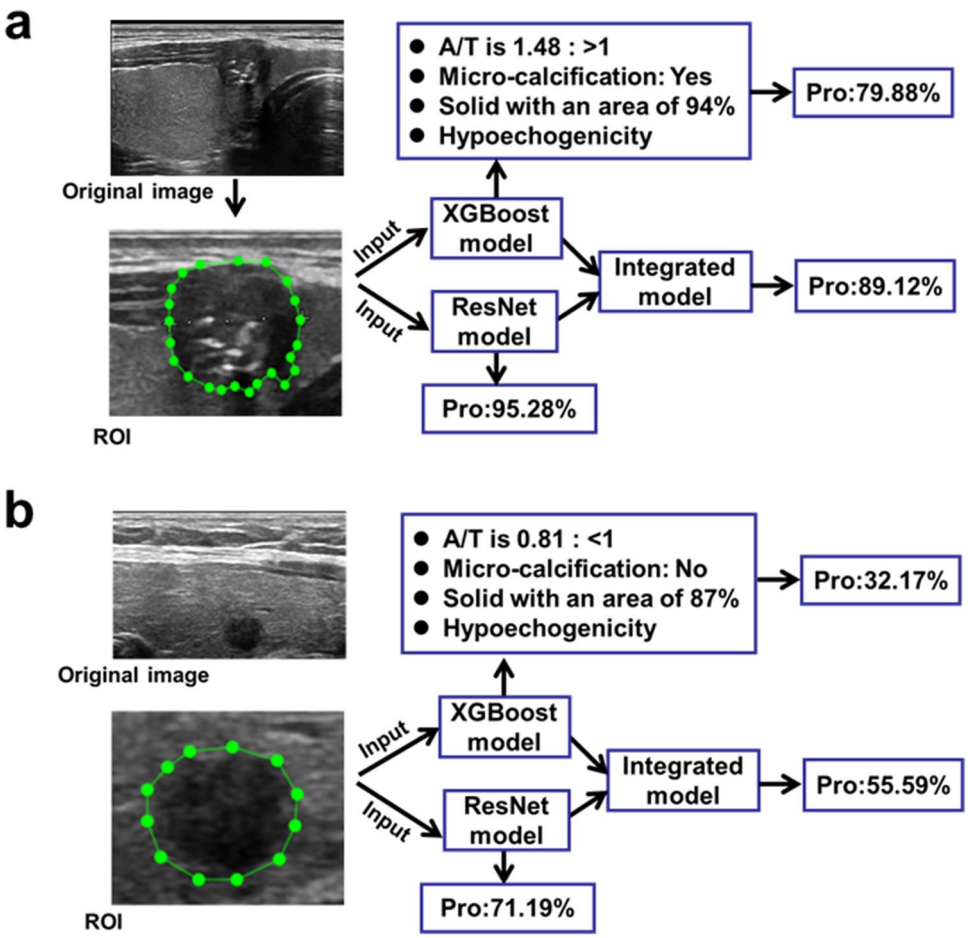

Fig. 4 Two representative cases for the real output of the integrated AI model. **a** A 44-year-old female suffering from papillary thyroid carcinoma with typical US risk features. The XGBoost model outputs the known US risk features, including solid composition with an area of 94%, positive microcalcifications, taller-than-wide shape (A/T = 1.48) and hypoechogenicity, and consequently gave a probability of 79.88% for malignancy. The ResNet model outputs a probability of 95.28% for malignancy, which was combined with the probability from the XGBoost model in an integrated model, leading to a final probability of 89.12%. **b** A 54-year-old female suffering from papillary thyroid carcinoma with atypical US risk features. The XGBoost model represented 87% solid and 13% cystic composition, negative microcalcification, hypoechogenicity, parallel shape (A/T = 0.81) and a probability of 32.17% for a malignancy. The ResNet model and the integrated model output risk probabilities of 71.19% and 55.59%, respectively. Pro, probability; ROI, region of interest; A/T, anteroposterior/transverse

Table 2). The AUC of the integrated model was significantly higher than that of the three junior sonographers ($p < 0.05$) but did not show a significant difference with the two seniors ($p > 0.05$) (Fig. 3, Table 2). The accuracy of the integrated model was significantly higher than that of two of the three junior sonographers (76.77% vs 61.62% and 65.66%, $p < 0.05$) but did not show a significant difference when compared to the two seniors (76.77% vs 73.23% and 72.72%, $p > 0.05$) (Table 2). The specificity of the integrated model was significantly higher than that of two of the three juniors (81.67% vs 67.50% and 42.50%, $p < 0.05$) and both seniors (81.67% vs 66.67% and 65.83%, $p < 0.05$) (Table 2).

Discussion

In this study, we retrospectively collected 2992 thyroid US images corresponding to 2992 solid or predominantly solid nodules, which were classified as TR 4 and TR 5 by two rich-experienced ultrasonographers adhering to ACR TI-RADS. We developed a new thyroid US CAD model by integrating two AI algorithms based on DL and traditional machine learning methods. The model not only shows high diagnostic accuracy for suspicious nodules but also displays the contribution of four known risk features to the prediction of thyroid cancer. Our model provides a problem-solving approach to ultrasonographers, thus

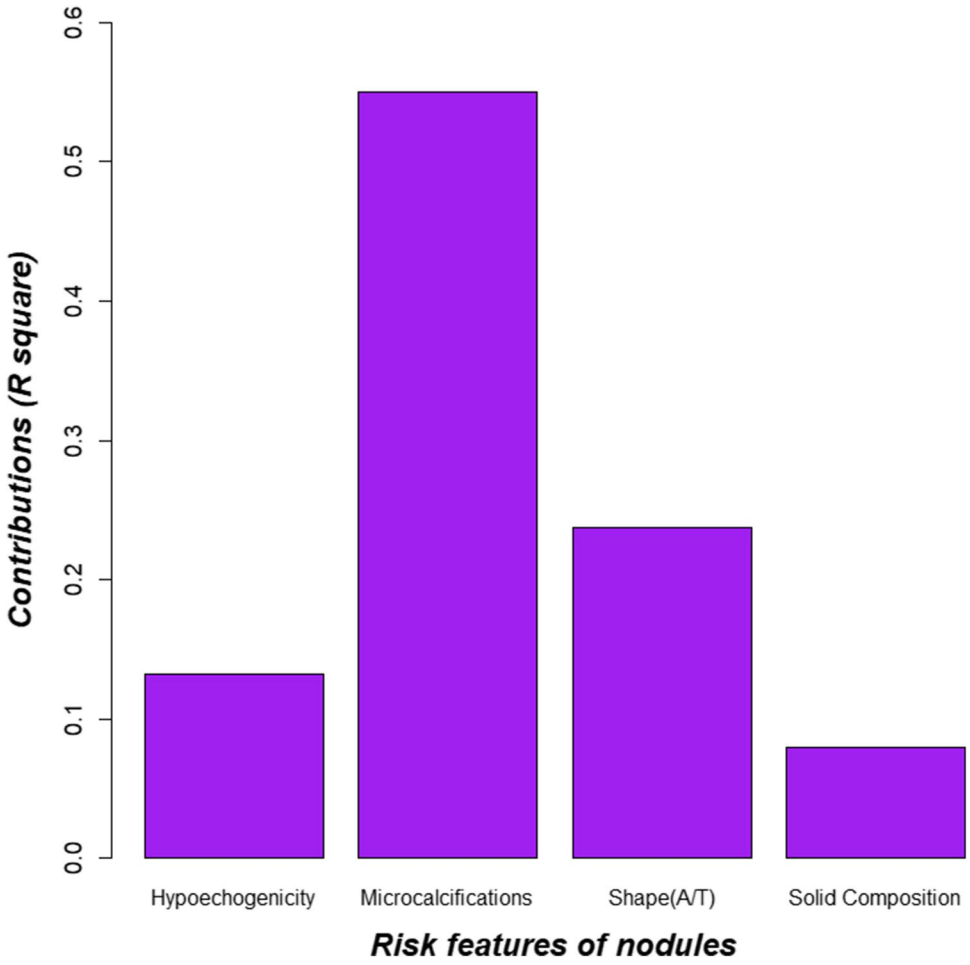

Fig. 5 Contribution of the risk feature of thyroid nodules to the malignancy diagnosis in the integrated model

promisingly improving the clinical acceptance of AI diagnosis of thyroid nodules.

ResNet50 is a deep learning method that has been widely used in medical image analysis. In this study, we adopted ResNet50 to train a thyroid US CAD model. The US CAD model showed a better performance than the five ultrasonographers (Table 2, Fig. 3). The result was consistent with previously reported thyroid US CAD models using similar deep learning methods [8–10]. In contrast to previous studies that covered all types of thyroid nodules ranging from TR1 to TR5, our study focused on analyzing solid or predominantly solid suspicious nodules classified as TR4 and TR5, which show much more complex features than TR1-3 nodules in clinical practice. Our US CAD model (accuracy: 75.25%) outperformed the young ultrasonographers (accuracy: 65.32%) for the complex TR4 and TR5 nodules, where some benign nodules showed echo features similar to those of thyroid cancers. Thus, our US CAD is a promising method for solving the actual difficulty in clinical diagnosis.

In addition to special attention to suspicious thyroid nodules, ultrasonographers also expect to obtain diagnostic evidence from the CAD system rather than the final result alone. DL-based methods work like "black boxes" and cannot explain nodule features. To challenge this issue, we trained another thyroid US CAD model using a traditional machine learning method named the XGBoost ensemble, by which the known features of malignant nodules were learned and interpreted automatically. The process was similar to the CAD S-Detect 2 software, which diagnosed a nodule as benign or malignant based on recognizing and assessing the features of the nodule, being able to obtain high sensitivity and better specificity than EU-TIRADS by ultrasonographers [18]. In contrast to the ResNet50 models, XGBoost took more time because of the additional procedure in which rich-experienced ultrasonographers label and define the features within thyroid nodules. However, the model has merit in that it can indicate the contribution of the risk features within a particular nodule to the diagnosis of malignancies. The results from the XGBoost-based US CAD model found that the microcalcifications showed the highest contribution with a weighting coefficient of 0.55, and the taller-than-wide shape showed the second-highest contribution with a coefficient of 0.24 (Fig. 5). Both findings are in

accordance with the clinical studies reported previously [19], which concluded that microcalcifications and taller-than-wide shapes are highly specific to malignant thyroid nodules. The hypoechogenicity and solid composition within thyroid nodules showed a relatively low contribution, consistent with known knowledge [20]. Compared with the deep-learning-based ResNet50 US CAD model, the XGBoost-based US CAD model showed higher specificity but lower sensitivity (Table 2), although both models achieved a better performance than the five ultrasonographers. The result indicates that the known risk features learned by the XGBoost model contributed to the diagnosis of malignant nodules, but there may exist some undiscovered risk features that were learned by the ResNet50 model but not recognized by ultrasonographers. These findings suggest that both models may compensate for each other, thus leading to the motivation to integrate them into one model.

By integrating the ResNet50 and XGBoost US CAD models, we achieved an improvement in diagnostic ability with an AUC of 0.80, a specificity of 81.67%, an accuracy of 76.77%, and a PPV of 71.05%, which were higher than those of both the ResNet50 and XGBoost models (Fig. 3, Table 2). In addition, the integrated model can display the known risk features similar to the XGBoost model. It can tell ultrasonographers what features were found within a particular nodule and how important those features are, thus also being a potential tool to train young ultrasonographers. Due to the higher accuracy in diagnosing suspicious thyroid nodules than humans and the explainability for the known features, the integrated model meets the standards for an ideal thyroid US CAD system to some extent and will improve the clinical acceptance of AI diagnosis of thyroid nodules.

However, there are still several limitations in this study, which need to be improved in the future. First, only static images were used in this study. The US image quality is often operator-dependent. A single scanning angle in a static image may cause the loss of important risk features for malignancy. Developing an AI model to evaluate cine clips may be a potential solution. Second, only the known risk features captured by the XGBoost model were explained in our integrated model. The risk features learned by the CNN model remain unexplained. The advancement of the methodology in the field of explainable AI will enlighten further exploration. Third, AI algorithms for risk features remain improvable, especially in the differentiation between microcalcifications and comet tails. Moreover, the value of the deep learning and traditional machine learning integrated model needs to be investigated further in a well-designed prospective study in multiple centers.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s00330-021-08298-7.

Acknowledgements We thank the ultrasonic physicians who took part in the labeling process of the study, as well as the engineers Yuli She and Yi Jing who gave good suggestions from a standpoint of medical image analysis. Last, we also want to give our sincere thanks to pathological physicians who offered the histopathological results for the thyroid nodules in our study.

Funding This study has received funding from the National Key Research and Development Program of China (No. 2017YFA0700800), the National Natural Science Foundation of China (No. 81871366), the Key Research and Development Program of Shaanxi Province of China (No. 2021SF-346).

Declarations

Guarantor The scientific guarantor of this publication is Qi Zhou.

Conflict of interest The authors of this manuscript declare no relationships with any companies whose products or services may be related to the subject matter of the article.

Statistics and biometry One of the authors has significant statistical expertise.

Informed consent Written informed consent was obtained from all subjects (patients) in this study.

Ethical approval Institutional Review Board approval was obtained.

Methodology
- Retrospective
- Diagnostic study
- Performed at one institution

References

1. Sung H, Ferlay J, Siegel RL et al (2021) Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin 71(3):209–249
2. Li M, Dal Maso L, Vaccarella S (2020) Global trends in thyroid cancer incidence and the impact of overdiagnosis. Lancet Diabetes Endocrinol 8(6):468–470
3. Durante C, Grani G, Lamartina L, Filetti S, Mandel SJ, Cooper DS (2018) The diagnosis and management of thyroid nodules: a review. JAMA 319(9):914–924
4. Itani M, Assaker R, Moshiri M, Dubinsky TJ, Dighe MK (2019) Inter-observer variability in the American College of Radiology Thyroid Imaging Reporting and Data System: in-depth analysis and areas for improvement. Ultrasound Med Biol 45(2):461–470
5. Vaccarella S, Franceschi S, Bray F, Wild CP, Plummer M, Dal Maso L (2016) Worldwide thyroid-cancer epidemic? The increasing impact of overdiagnosis. N Engl J Med 375(7):614–617
6. Tessler FN, Middleton WD, Grant EG et al (2017) ACR Thyroid Imaging, Reporting and Data System (TI-RADS): white paper of the ACR TI-RADS Committee. J Am Coll Radiol 14(5):587–595
7. Hoang JK, Middleton WD, Farjat AE et al (2018) Reduction in thyroid nodule biopsies and improved accuracy with American College of Radiology Thyroid Imaging Reporting and Data System. Radiology 287(1):185–193

8. Buda M, Wildman-Tobriner B, Hoang JK et al (2019) Management of thyroid nodules seen on US images: deep learning may match performance of radiologists. Radiology 292(3):695–701
9. Thomas J, Ledger GA, Mamillapalli CK (2020) Use of artificial intelligence and machine learning for estimating malignancy risk of thyroid nodules. Curr Opin Endocrinol Diabetes Obes 27(5):345–350
10. Thomas J, Haertling T (2020) AIBx, Artificial intelligence model to risk stratify thyroid nodules. Thyroid 30(6):878–884
11. Ye FY, Lyu GR, Li SQ et al (2021) Diagnostic performance of ultrasound computer-aided diagnosis software compared with that of radiologists with different levels of expertise for thyroid malignancy: a multicenter prospective study. Ultrasound Med Biol 47(1):114–124
12. Morid MA, Borjali A, Del Fiol G (2021) A scoping review of transfer learning research on medical image analysis using ImageNet. Comput Biol Med 128:104115
13. Liu SF, Wang Y, Yang X (2019) Deep learning in medical ultrasound analysis: a review. Engineering 5:261–275
14. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA. Available via https://ieeexplore.ieee.org/document/7780459. Accessed 16 Aug 2021
15. Krizhevsky A, Sutskever I, Hinton G (2017) ImageNet classification with deep convolutional neural networks. Commun ACM 60(6):84–90
16. Shin JH, Baek JH, Chung J et al (2016) Ultrasonography diagnosis and imaging-based management of thyroid nodules: revised Korean Society of Thyroid Radiology Consensus Statement and Recommendations. Korean J Radiol 17(3):370–395
17. Torlay L, Perrone-Bertolotti M, Thomas E, Baciu M (2017) Machine learning-XGBoost analysis of language networks to classify patients with epilepsy. Brain Inform 4(3):159–169
18. Szczepanek-Parulska E, Wolinski K, Dobruch-Sobczak K et al (2020) S-Detect software vs. EU-TIRADS classification: a dual-center validation of diagnostic performance in differentiation of thyroid nodules. J Clin Med 9(8):2495
19. Remonti LR, Kramer CK, Leitão CB, Pinto LC, Gross JL (2015) Thyroid ultrasound features and risk of carcinoma: a systematic review and meta-analysis of observational studies. Thyroid 25(5):538–550
20. Rosario PW, Mourão GF (2019) Noninvasive follicular thyroid neoplasm with papillary-like nuclear features (NIFTP): a review for clinicians. Endocr Relat Cancer 26(5):R259–R266